IDIAP RESEARCH REPORT

COMPARATIVE STUDY ON SENTENCE BOUNDARY PREDICTION FOR GERMAN AND ENGLISH BROADCAST NEWS

Yang Wang [a] Alexandre Nanchen
Alexandros Lazaridis [b] David Imseng
Philip N. Garner

Idiap-RR-18-2017

JULY 2017

[a] Idiap
[b] Idiap Research Institute

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Comparative Study on Sentence Boundary Prediction for German and English Broadcast News

Yang Wang, *Alexandre Nanchen*, *Alexandros Lazaridis*, *David Imseng, Philip N. Garner*

Idiap Research Institute, Martigny, Switzerland

{yang.wang,alexandre.nanchen,alaza,dimseng,phil.garner}@idiap.ch

Abstract

We present a comparative study on sentence boundary prediction for German and English broadcast news that explores generalization across different languages. In the feature extraction stage, word pause duration is firstly extracted from word aligned speech, and forward and backward language models are utilized to extract textual features. Then a gradient boosted machine is optimized by grid search to map these features to punctuation marks. Experimental results confirm that word pause duration is a simple yet effective feature to predict whether there is a sentence boundary after that word. We found that Bayes risk derived from pause duration distributions of sentence boundary words and non-boundary words is an effective measure to assess the inherent difficulty of sentence boundary prediction. The proposed method achieved F-measures of over 90% on reference text and around 90% on ASR transcript for both German broadcast news corpus and English multi-genre broadcast news corpus. This demonstrates the state of the art performance of the proposed method.

Index Terms: sentence boundary prediction, punctuation, broadcast news, word pause duration, gradient boosted machine

1. Introduction

Transcripts outputted by Automatic Speech Recognition (ASR) systems are typically a sequence of words without sentence boundaries, punctuation marks or case differentiation. However, the lost information is important to readability of ASR transcripts for both human understanding and for further natural language processing tasks. Of these three sources of information, sentence boundaries play a key role in readability since they segment sequences of words into more meaningful sentences. In this comparative study, we investigate sentence boundary prediction for German and English broadcast news corpora, aiming at exploring the possibility of language independence. We address three research questions:

1. Does sentence boundary prediction accuracy mainly depend upon language?
2. What is the relative importance of acoustic and language model cues?
3. What is the best possible performance given the proposed method in this study?

From an acoustic feature viewpoint, we investigate word pause duration distributions for sentence boundary words and non-boundary words, and calculate Bayes risk empirically based on these distributions, confirming that Bayes risk in this scenario is an intuitive, simple and effective measure to assess the inherent difficulty of sentence boundary prediction. From a textual feature viewpoint, both forward and backward *n*-gram Language Models (LMs) for sentence boundary and clause boundary are utilized. Then a Gradient Boosted Machine (GBM) [1] combines all these features to produce a final prediction. Experimental results show that the proposed method produces the state of the art performance, and smaller empirical Bayes risk indeed corresponds to better sentence boundary prediction performance.

2. Related work

Sentence boundary prediction and punctuation prediction have been explored over two decades. Stolcke and Shriberg [2, 3] firstly used an *n*-gram LM, speaker turn information, and Part-of-Speech (POS) to predict sentence boundaries for ASR transcripts, later they presented a similar method to detect both sentence boundaries and disfluencies using more features, such as duration, pitch, and energy [4]. Chen reported a method of predicting various types of punctuation marks by regarding punctuation marks as special words, in which a slightly extended LM was utilized to accommodate punctuation and decode in a uniform way [5].

Since then a variety of approaches to this problem emerged, which may be briefly summarized from the aspects of features and modelling methods as follows:

2.1. Features

In most research work either only textual features or combined textual features and acoustic features are used. While the most commonly used textual feature is *n*-gram statistical LM [2-4, 6], POS [7], syntactical information [8, 9], and dysfluency annotation [10, 11] have also been investigated. Regarding acoustic features, many prosodic features are demonstrated to be useful, such as pause, word or phoneme duration, energy, fundamental frequency, and heuristically derived prosodic features [3, 12-14].

2.2. Methods

Three main types of methods have been developed for sentence boundary prediction and punctuation prediction. The first type is to regard punctuation as special words between normal words, usually predicted by an extended LM which is trained on a text corpus containing punctuation. The second type is to consider this problem as a word labeling problem, in which each word is assigned to a punctuation mark or no punctuation. Many methods from the machine learning community can be used, such as maximum entropy [15], boosting [16], conditional random fields [8]. The third, more recent method is to treat it as a mono-lingual machine translation problem which translates word sequence without punctuations to segmented and punctuated text [17, 18].

3. Proposed method

We developed a combined acoustic and textual method for sentence boundary prediction. Since extracted features are themselves weak classifiers, these features are further combined in a GBM to produce the final prediction. The architecture of the proposed method is shown in Figure 1, in which the left column corresponds to the acoustic feature extraction, the right column explains the textual feature extraction, and the bottom part shows combination in GBM. We describe this method in more detail below.

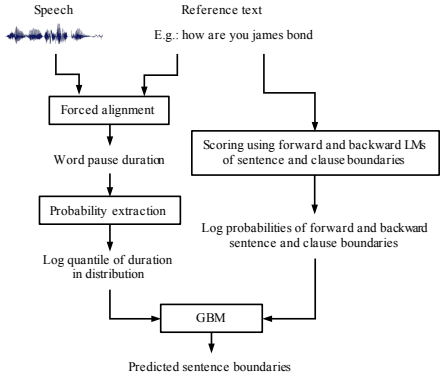

Figure 1: *Architecture of proposed method.*

3.1. Feature extraction

3.1.1. Acoustic feature: word pause duration

From aligned speech with reference text or decoded speech with ASR transcripts, it is trivial to calculate a pause duration after each word. We will call it pause duration or duration hereafter. Typically, sentence boundary words have longer pause duration, and non-boundary words have zero or shorter pause duration. It is thus widely accepted that this is a useful feature to predict sentence boundary, e.g. [12, 19]. However, pause durations obtained by an automatic alignment tool can contain excessively large values (10 seconds) occasionally, especially when aligning erroneous reference text or very noisy ASR transcripts.

This study does not use the pause duration value directly, but uses the quantile of pause duration in the pause duration distribution of all words in a corpus. The benefit of this replacement is: it is more robust to outliers, and quantile values are easily used as a probabilistic measure in the following processing steps. This one-dimensional feature of pause duration quantile in log scale is used as input to a GBM.

3.1.2. Textual features: LM based boundary probabilities

An *n*-gram LM is typically used as baseline system. Instead of the common setup of using only one LM, we tried to use four LMs to extract more textual information.

Firstly, in addition to the commonly used LM for predicting the possibility of *n*-gram followed by a sentence boundary (*forward LM*), we add another independent *backward LM* to predict the possibility of an *n*-gram preceded by a sentence boundary. Secondly, our preliminary experiment shows that sentence boundaries are correlated with clause boundaries indicated by commas, and less correlated with normal words without any punctuation marks. This phenomenon is in accord with our feeling that clause boundaries are sometimes quite subjective, and could be replaced by sentence boundaries. E.g., in the reference text "Hello, James, how are you?", the second comma can be replaced by a full stop arbitrarily. In short, sentence boundary is correlated with clause boundary, thus clause boundary can contribute to sentence boundary prediction. As a result, we also build forward and backward LMs of clause boundaries for sentence boundary prediction. These LMs leads to four probabilities extracted as textual features.

3.2. Classifier

The acoustic and textual features were combined using a GBM to predict sentence boundaries. We chose GBM as it is a widely used learning machine for its robust performance and ability to perform well with outliers. A broad grid search over hyper parameters is done to train the GBM.

4. Corpus and analysis

This section firstly presents an overview of the read German BCN corpus [20] and the spontaneous English MGB corpus [21], then explains intuitively the utility of pause duration from its distributions, and calculates Bayes risks derived from above distributions. Then we tried to optimize *n* in *n*-gram LM in terms of perplexity. Finally, a GBM is used as a predictor.

4.1. Corpora overview

The German BCN corpus consists of over 160 hours of high quality, manually segmented and annotated radio broadcast news, most of which is clearly pronounced like read speech. Component utterances are typically 1 to 3 minutes; each utterance is typically composed of several sentences. The only type of punctuation mark in this corpus is full stop, indicating sentence boundaries. Thus, this corpus is ideal for our experiment, since it is clean, and most utterances are of reasonable length with several sentence boundaries to be predicted. To train German LMs, we used the German text in the Europarl corpus [22] and the training text in the BCN corpus.

The MGB corpus consists of 1600 hours of recorded television programs. Since the transcripts were obtained from broadcast subtitles, they contain ASR errors, thus the Matching Error Rate (MER) is also supplied to indicate how well the transcript corresponds to its audio given a trained acoustic model. To reduce the effect of high Word Error Rate (WER) of around 30% in ASR transcripts, a subset of 240 hours with MER less than 10% was selected from training data. However, evaluation data were not selected in a similar way since we also want to predict sentence boundary for its noisy ASR transcripts. The extracted utterances are typically short, consisting of only one or two sentences, thus fewer sentence boundaries need to be predicted compared with the BCN corpus. This makes a significant difference when explaining experimental result. The MGB corpus provides a large text for LM training as well. This corpus was firstly preprocessed at CSTR [17], having five types of punctuation marks; these need to be normalized further to match a common utterance pattern assumption below.

4.1.1. Utterance pattern assumption

We assume that each utterance is a sequence of sentences, and each sentence is a sequence of words followed by a sentence boundary. It is also possible that one utterance has only one

sentence. This utterance pattern assumption naturally holds for the BCN corpus, and holds for most but not all utterances in the MGB corpus, thus the following normalization described in next section is used to make sure that the assumption is hold for fair comparison over the two corpora.

It should be noted that this assumption itself will give some prior sentence boundary prediction, since there is always a sentence boundary at the end of each utterance even if we do not use any acoustic and textual features.

4.1.2. Text normalization for the MGB corpus

The normalization rules used for the MGB corpus are shown in Table 1, in which the more complex cases in the last two rows were removed for simplicity in this study.

Table 1: *Normalization rules for the MGB corpus*

Cases	**Rules**
Full stop, question mark, exclamation mark, three dots	Convert to sentence boundary
Comma	Remove comma punctuation
Multiple punctuation (e.g., !?)	Remove whole utterance
Utterance without punctuation	Remove whole utterance

4.1.3. Basic statistics of two corpora

A few basic statistics of the two corpora are shown in Table 2, in which "utt." and "sent." are abbreviations for utterances and sentences, respectively.

Table 2: *Basic statistics of the two corpora.*

Items	**Statistics**			
	German BCN		English MGB	
	train	test	train	test
# utt.	12766	1265	143947	8320
# sent.	74567	6821	206378	9040
# words	1013996	100615	1886647	54105
# sent. / # utt.	5.84	**5.39**	1.43	**1.09**

4.1.4. Sentence length distribution

The distribution of sentence length in terms of number of words in a corpus is an intuitive way to understand how sentence boundaries are distributed among word sequences. The sentence length distributions of these two corpora are shown in Figure 2.

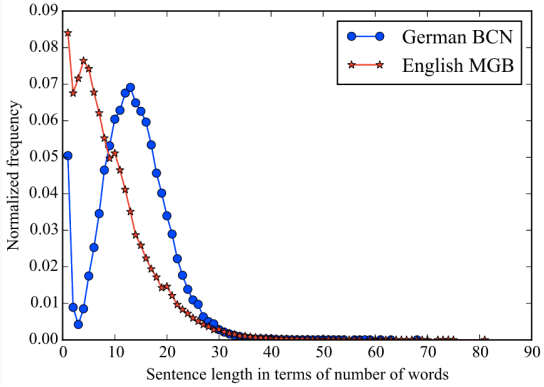

Figure 2: *Sentence length distribution.*

From this figure and the corpus text, we can see that BCN sentences have one short sentence pattern that usually contains only a few words ("Berlin."), and one long sentence pattern being normal sentences. By contrast, MGB sentences have a monotonically decreasing trend where longer sentences are less likely to appear.

4.2. Feature analysis

4.2.1. Word pause duration

As discussed in section 3.1.1, pause duration can be used to classify boundary words and non-boundary words, as shown in Figure 3 for the BCN corpus and in Figure 4 for the MGB corpus.

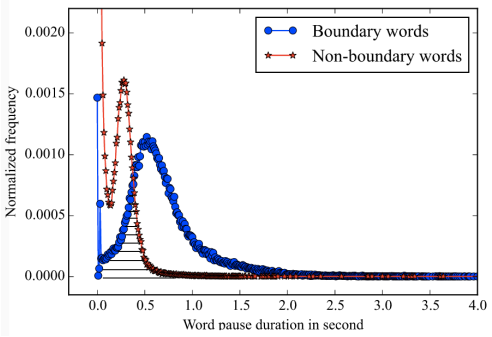

Figure 3: Pause duration distribution on BCN corpus.

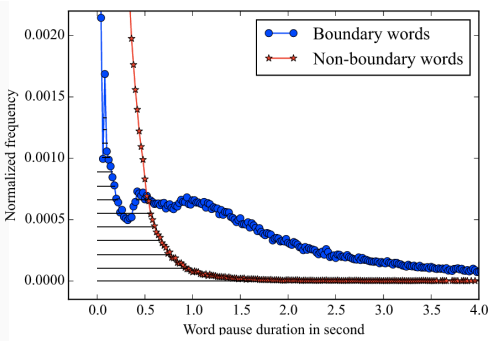

Figure 4: Pause duration distribution on MGB corpus.

From Figure 3, we can see that, statistically, German boundary words indeed have longer pause durations compared with non-boundary words. Further, non-boundary words have a clear peak at about 300ms; we assume that this peak corresponds to clause boundary. Although this assumption cannot be verified directly since no clause boundaries are annotated, listening to the speech files reveals that this assumption is reasonable.

From Figure 4, we know that English words have similar trends except for two exceptions. The pause duration peak hypothesized for clause boundary words is completely missing, and this is confirmed in the pause duration feature file where many commas indeed have zero or quite short pause duration. Another exception is that a noticeable portion of boundary words has very short pause duration. These phenomena probably come from the fact that the MGB corpus is spontaneous, thus the pause durations of its boundary words are inherently more irregular and more difficult to predict.

The separability of boundary words and non-boundary words when using only pause duration can be quantified by measuring the overlapped/shaded area of two distributions in each figure, which is the minimum Bayes risk in theory. This

risk is calculated empirically here since all pause durations and frequencies are discrete, thus it is easy to accumulate their overlapped area. This lower bound risk is denoted as $P(\text{bound}|\text{dur})$, i.e., predicting sentence boundary using posterior pause duration information. On the other hand, the prior probability of whether a word is a sentence boundary word gives an upper bound of the classification risk, denoted by $P(\text{bound})$, i.e., predicting sentence boundary without any posterior information. These risks are shown in Table 3. Clearly, the Bayes risk of the BCN corpus is much smaller than that of the MGB corpus, indicating better separability of boundary and non-boundary, thus the BCN corpus is expected to have better prediction performance than the MGB corpus.

Table 3: *Classification risks using pause duration*

Risks	**German BCN** (%)	**English MGB** (%)
$P(\text{bound})$	7.30	10.93
$P(\text{bound}\|\text{dur})$	**1.97**	**4.95**

4.2.2. LM

We chose 3-gram LM for the BCN corpus, and 4-gram LM for the MGB corpus based on LM's perplexity.

5. Experiments

The above acoustic and textual features are input to a GBM tool [23], and the GBM's hyper-parameters are optimized by a wide grid search. The overall performance of sentence boundary prediction is shown in Table 4 and Table 5, in which "UP" means the "a-priori" performance by just using the Utterance Pattern (UP) assumption discussed in section 4.1.1, "Dur" means pause duration, and P., R., and F. mean precision, recall, and F-measure, respectively.

We see that for the BCN corpus, pause duration contributes more than the LM and UP assumption. When combined, all three features together produce the best performance, and the UP assumption play a negligible role in improving performance when duration and LM features are already used. Notice that when only the UP assumption is used, the precision is 100% theoretically, and the recall rate (18.55%) is the number of utterances divided by the number of sentences, i.e., the inverse of the averaged utterance length in terms of number of sentences, as shown in last row of Table 2 (i.e., 5.39).

Table 4: *Results on the BCN corpus.*

Features	**Reference text (%)**			**ASR transcript (%)**		
	P.	R.	F.	P.	R.	F.
UP	100.00	18.55	**31.29**	100.00	18.55	**31.29**
Dur	83.75	86.30	85.01	83.01	87.69	85.29
LM	78.74	58.06	66.84	77.75	56.82	65.66
Dur+LM	91.47	88.99	90.21	91.35	88.12	89.71
Dur+UP	87.68	87.04	87.36	88.11	86.93	87.52
LM+UP	82.52	64.33	72.30	81.64	63.20	71.25
Dur+LM+UP	91.91	89.25	**90.56**	91.67	88.18	**89.89**

However, for the English MGB corpus, the UP assumption dominates the final F-measure by a very high a-priori F-measure of 95.32%; adding pause duration or textual features typically degrades performance. The high a-priori F-measure comes from the fact that the MGB corpus is very fragmented, and on average each utterance has only 1.09 sentences, leading to 1/1.09≈92% recall rate and almost 100% precision.

To reduce the high prior of F-measure, a subset of long utterances containing at least two sentence boundaries were selected to do a similar analysis, as shown in Table 6. Now we can see that the LM contributes more than the duration feature and UP assumption, and the best F-measure was achieved when combining all features and UP assumption.

Considering German and English ASR transcripts have WERs of 6.5% and 34.2%, respectively, the degradation of sentence boundary prediction when applied to ASR transcripts instead of reference text is small, e.g., from 81.94% to 76.05% in Table 6, compared to the result of a similar task from 87.03% to 74.41% in Table 5 of [17]. This is because both the pause duration feature and limited context *n*-gram LMs are robust to ASR errors, and the GBM itself is robust as well.

Table 5: *Results on the MGB corpus.*

Features	**Reference text**			**ASR transcript**		
	P.	R.	F.	P.	R.	F.
UP	99.44	91.52	**95.32**	99.43	91.38	**95.24**
Dur	97.78	44.37	61.04	96.66	45.92	62.26
LM	76.21	50.79	60.96	72.02	40.22	51.62
Dur+LM	93.27	68.08	78.71	93.19	62.40	74.75
Dur+UP	99.44	91.52	95.32	99.43	91.38	95.24
LM+UP	96.23	93.84	95.02	96.43	92.73	94.54
Dur+LM+UP	96.22	94.02	**95.11**	96.59	92.99	**94.76**

Table 6: *Results on a subset of the MGB corpus containing at least two sentence boundaries.*

Features	**Reference text**			**ASR transcript**		
	P.	R.	F.	P.	R.	F.
UP	99.11	46.48	**63.28**	99.38	49.57	**66.15**
Dur	95.67	27.77	43.05	91.95	27.38	42.20
LM	77.32	62.11	68.89	72.66	48.11	57.89
Dur+LM	81.03	75.44	78.14	77.48	62.26	69.04
Dur+UP	99.11	46.48	63.28	99.38	49.57	66.15
LM+UP	86.42	72.37	78.77	86.16	66.90	75.32
Dur+LM+UP	82.52	81.37	**81.94**	80.19	72.31	**76.05**

6. Conclusions

We can now try to answer the three research questions:

1. Although our results appear to show similar performance for both German and English, it is difficult to draw a strong conclusion as the corpora also differ significantly in speaking style. Speaking style may play a more important role than language for F-measure in this study.
2. Duration is more important than LM for the read German BCN corpus, but less important for the spontaneous English MGB corpus. The UP assumption dominates F-measure for the MGB corpus, but not for the BCN corpus.
3. Sentence boundaries can be predicted about 90% in terms of F-measure for both languages, and 81.94% for reference text in subset of MGB corpus in Table 6. Although working on different corpus, this is roughly comparative to the state of the art (81.0% in [24] and 84.1% in [25] measured by slightly modified F1 measure).

7. Acknowledgements

This paper was support by the H2020 project SUMMA. Authors would like to express thanks to Ondrej Klejch for providing preprocessed MGB database to us.

8. References

[1] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics,* pp. 1189-1232, 2001.

[2] A. Stolcke and E. Shriberg, "Automatic linguistic segmentation of conversational speech," in *ICSLP* 1996, pp. 1005-1008.

[3] A. Stolcke, "Modeling Linguistic Segment and Turn Boundaries for N-best Rescoring of Spontaneous Speech," in *EUROSPEECH*, 1997.

[4] A. Stolcke *et al.*, "Automatic detection of sentence boundaries and disfluencies based on recognized words," in *ICSLP*, 1998, pp. 2247-2250.

[5] C. J. Chen, "Speech Recognition with Automatic Punctuation," in *Eurospeech*, 1999.

[6] A. Gravano, M. Jansche, and M. Bacchiani, "Restoring Punctuation and Capitalization in Transcribed Speech," in *ICASSP*, 2009, pp. 4741-4744.

[7] E. Cho, K. Kilgour, J. Niehues, and A. Waibel, "Combination of NN and CRF models for joint detection of punctuation and disfluencies," in *Interspeech*, 2015, pp. 3650-3654.

[8] U. Nicola, B. Maximilian, and V. Paul, "Improved models for automatic punctuation prediction for spoken and written text," in *Interspeech*, 2013.

[9] D. Zhang, S. Wu, N. Yang, and M. Li, "Punctuation Prediction with Transition-based Parsing," in *ACL*, 2013, pp. 752-760.

[10] M. Meteer and R. Iyer, "Modeling conversational speech for speech recognition," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1996, pp. 33-47.

[11] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching Speech Recognition With Automatic Detection of Sentence Boundaries and Disfluencies," *IEEE Transactions on audio, speech, and language processing,* vol. 14, no. 5, pp. 1526-1540, 2006.

[12] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," in *Proceedings of the International Speech Communication Association (ISCA) Workshop: Automatic Speech Recognition: Challenges for the New Millennium (ASR-2000)*, Paris, 2000.

[13] J.-H. Kim and P. C. Woodland, "The Use of Prosody in a Combined System for Punctuation Generation and Speech Recognition," in *Eurospeech*, 2001.

[14] M. Zimmerman *et al.*, "The ICSI+ Multilingual Sentence Segmentation System," DTIC Document2006.

[15] J. Huang and G. Zweig, "Maximum Entropy Model for Punctuation Annotation from Speech," in *Interspeech*, 2002.

[16] J. Kolár and L. Lamel, "Development and Evaluation of Automatic Punctuation for French and English Speech-to-Text," in *Interspeech*, 2012, pp. 1376-1379.

[17] O. Klejch, P. Bell, and S. Renals, "Punctuated Transcription of Multi-genre Broadcasts Using Acoustic and Lexical Approaches," in *IEEE Workshop on Spoken Language Technology*, 2016.

[18] O. Klejch, P. Bell, and S. Renals, "Sequence-to-Sequence Models for Punctuated Transcription Combing Lexical and Acoustic Features," in *ICASSP 2017*, pp. 5700-5704.

[19] G. Dzhambazov and R. Bardeli, "Automatic Sentence Boundary Detection for German Broadcast News," in *Speech Communication; 10. ITG Symposium; Proceedings of*, 2012, pp. 1-4: VDE.

[20] F. Weninger, B. Schuller, F. Eyben, M. Wöllmer, and G. Rigoll, "A Broadcast News Corpus for Evaluation and Tuning of German LVCSR Systems," 2014.

[21] P. Bell *et al.*, "The MGB Challenge: Evaluating multi-genre broadcast media recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, 2015, pp. 687-693: IEEE.

[22] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT summit*, 2005, vol. 5, pp. 79-86.

[23] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research,* vol. 12, no. Oct, pp. 2825-2830, 2011.

[24] C. Xu, L. Xie, G. Huang, X. Xiao, E. Chng, and H. Li, "A Deep Neural Network Approach for Sentence Boundary Detection in Broadcast News," in *Interspeech*, 2014, pp. 2887-2891.

[25] G. Huang, C. Xu, X. Xiao, L. Xie, E. S. Chng, and H. Li, "Multi-View Features in a DNN-CRF Model for Improved Sentence Unit Detection on English Broadcast News," in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, 2014, pp. 1-9: IEEE.